

Exploration of Proximity Heuristics in Length Normalization

Pranav A

VIT University

Vellore, India

cs.pranav.a@gmail.com

Abstract

Ranking functions used in information retrieval are primarily used in the search engines and they are often adopted for various language processing applications. However, features used in the construction of ranking functions should be analyzed before applying it on a data set. This paper gives guidelines on construction of generalized ranking functions with application-dependent features. The paper prescribes a specific case of a generalized function for recommendation system using feature engineering guidelines on the given data set. The behavior of both generalized and specific functions are studied and implemented on the unstructured textual data. The proximity feature based ranking function has outperformed by 52 % from regular BM25.

1 Introduction

Information retrieval deals with the task of *retrieving*, or in simple words, obtaining relevant and necessary information resources from a huge collection of documents. The information obtained is considered relevant according to a piece of information asked about (also known as *query*) (Zhai and Massung, 2016). The use of information retrieval has been fundamental for developing search engines. But there are many other interesting applications such as recommendation systems, spam filtering, plagiarism detection and so on (Manning et al., 2008). Ranked information retrieval use ranking functions which determine the decreasing order of the documents in relevance to the query.

Zhai and Massung (2016) lay down the in-depth analysis of variety of ranking functions used in information retrieval. Common features used in

the information retrieval are term frequency, inverse document frequency and length normalization. TF-IDF (Term Frequency and Inverse Document Frequency) is a prevalent vector-space information retrieval technique. It balances the similarity of the query and the document, and penalizes the common terms (Wu et al., 2008). Pivoted document length normalization with TF-IDF helps to reward shorter documents (Singhal et al., 1996). Okapi BM25 is a complex version of pivoted length normalization and it is the prevailing state-of-the-art retrieval method (Robertson and Walker, 1994). PL2 is a retrieval model based on divergence from randomness of the query term frequency (Amati and Van Rijsbergen, 2002). Dirichlet Prior based retrieval is based on language modeling where the smoothing function is derived from Dirichlet distribution (Zhai and Lafferty, 2004). MPtf2ln and MDtf2ln are improvements on previous methods and more elaborate ranking functions which balances the extremities of normalization effects (Fang et al., 2011). Fang et al. (2011) also lay down certain guidelines to evaluate the behavior of the ranking functions.

This paper analyses the feature engineering concepts in the information retrieval, especially length normalization. The objective of this paper are:

- to give general guidelines of general structure and nature of feature which can be included in the ranking function. (Section 2)
- to study the existing work which use information retrieval in recommendation systems. (Section 3)
- to prepare an unstructured textual dataset for penpal recommendation system. (Section 4)
- to analyze the dataset and apply the general

guidelines to construct the feature for penpal recommendation system. (Section 5)

- to test and compare the constructed function against other ranking functions. (Section 6)

2 Guidelines for Feature Engineering in Ranking Functions

The main idea of the information retrieval systems is to include a proximity score in conjunction to the ranking function. A proximity score is an application-dependent score, like page-rank metrics, contextual similarity and so on. The function should have well engineered feature embedded in itself and it should be able to generalize other results. The objectives of the ranking function are:

- The score should be higher if the constituents of proximity score (real numbers or vectors) are similar.
- The proximity score should be below certain cut-off scores. These cut-off scores or limiting parameters can be determined through machine learning.
- The function should behave similar and able to generalize the behavior and working of the existing ranking functions in IR.

The assumption is taken that modeled function is bi-modal (as one has to consider extremities of the cut-off scores). Let $f(\mathbf{x}, \mathbf{y})$ be the modeled function, where \mathbf{x} and \mathbf{y} be the constituent vectors. In other words, this function determines similarity between \mathbf{x} and \mathbf{y} and it is a function of $d(\mathbf{x})$ and $d(\mathbf{y})$ where $d(\cdot)$ is the distance metric.

Length normalization functions, especially in the BM25 or pivoted length function are inversely proportional to the to the TF-IDF score. Thus, the model would be $f(\mathbf{x}, \mathbf{y})$ inversely proportional to the TF-IDF score.

$$TF - IDF(q, d) \propto \frac{1}{f(\mathbf{x}, \mathbf{y})}$$

2.1 Nature of the Function Curve

Since the function has to be bi-modal distribution, distribution function would be modeled in the form asymmetrical inverted bell curve. Thus the trough of the curve should occur when distances of \mathbf{x} and \mathbf{y} are similar. In other words,

$$\frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial(d(\mathbf{x}))} = 0 \text{ if } d(\mathbf{x}) \approx d(\mathbf{y}) \quad (1)$$

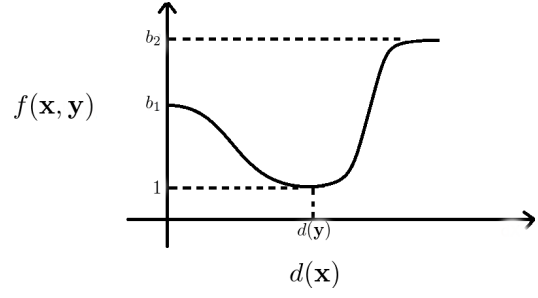


Figure 1: Rough Sketch of Feature Curve

Similarly it can be shown that,

$$\frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial(d(\mathbf{y}))} = 0 \text{ if } d(\mathbf{x}) \approx d(\mathbf{y})$$

From here onwards, arguments would be taken from the $d(\mathbf{x})$ perspective, as $d(\mathbf{y})$ would have similar arguments.

The function should decrease monotonically if $d(\mathbf{x}) < d(\mathbf{y})$. In other words,

$$\frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial(d(\mathbf{x}))} < 0 \text{ if } d(\mathbf{x}) < d(\mathbf{y}) \quad (2)$$

The function should increase monotonically if $d(\mathbf{x}) > d(\mathbf{y})$. In other words,

$$\frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial(d(\mathbf{x}))} > 0 \text{ if } d(\mathbf{x}) > d(\mathbf{y}) \quad (3)$$

2.2 Nature of the Limits of the Curve

Here the left extremity of the curve could be bounded to the parameter b_1 .

$$\lim_{d(\mathbf{x}) \rightarrow 0} f(\mathbf{x}, \mathbf{y}) = b_1 \quad (4)$$

Similarly, the right extremity of the curve could be bounded to the parameter b_2 .

$$\lim_{d(\mathbf{x}) \rightarrow \infty} f(\mathbf{x}, \mathbf{y}) = b_2 \quad (5)$$

The trough of the curve exists when $\frac{\partial f(\mathbf{x}, \mathbf{y})}{\partial(d(\mathbf{x}))} = 0$. Thus the limit is set to 1. It can not be set it to 0, because $f(\mathbf{x}, \mathbf{y})$ lies in the denominator of the scoring function.

$$\lim_{d(\mathbf{x}) \rightarrow d(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) = 1 \quad (6)$$

The specific case of these guidelines can be applied to the recommendation systems.

3 Information Retrieval in Recommendation Systems

Not many efforts have been carried out which use ranking functions in recommendation systems. Most of the recommendation systems use latent Dirichlet allocation (LDA) or collaborative filtering. Adomavicius and Tuzhilin (2005) give a more comprehensive study of existing research work on recommendation systems.

Some recommendation systems have experimented with ranking functions. However, all of them did not tune the parameters correctly. The systems which deploy IR functions, have generally used extremely well-structured data like tags (Wang and Yuan, 2009) or contextual variables (Kwon and Kim, 2010).

Tag-based systems : These systems have concatenated the tags of an user into a single *document*. Relevant proximity is determined through similarity between *query* and tags. This idea has used to develop music-recommendation systems (Cantador et al., 2010; Bellogín et al., 2010). The cosine similarity between BM25 value of user profile tags and item profile tags produced the best results. This leads to conclusion that BM25 outperforms other functions because BM25 penalizes common tags and focuses on rare terms (Cantador et al., 2010). The effects of length normalization are not discussed. More evaluation metrics have been introduced to evaluate the ranking functions (Bellogín et al., 2010). Results concluded that collaborative filtering methods have more novel and diverse recommendations and ranking functions have more coverage and better accuracy. TF-IDF outperforms BM25 but the parameters of BM25 have not been tuned.

Publication-recommendation systems : Some publication-recommendation systems have used ranking functions (Nishioka and Scherp, 2016). Concept Frequency Inverse Document Frequency (CF-IDF) and Hierarchical CF-IDF (HCF-IDF) have been utilized which generally assign weights on certain pre-determined words in the documents. The results showed that CF-IDF combined with sliding window produced the best results. Surprisingly, it outperforms the popular LDA method. Another publication-recommender system have used information retrieval elements (Totti et al., 2016). Here citation context (TF-IDF similarity between two papers), query similarity (TF-IDF similarity between query and arti-

cle) and age decay (to penalize older articles) have been considered as parameters. This experiment also showed that it surpasses the system that uses page-rank like metrics.

Recommendation systems that use unstructured data: Suchal and Návrat (2010) and Esparza et al. (2012) have used unstructured textual data in the recommendation system designs. Users' choices have been concatenated into a single query which is used to recommend articles (Suchal and Návrat, 2010). Movie recommendation has also been evaluated with ranking functions (Esparza et al., 2012). Tags and reviews of movies are served as an TF-IDF component. Reviews tend to have better performance than tags because unstructured nature of reviews provide some noise and undiscovered information as opposed to structured data like tags. This results high IDF values. Thus unstructured textual data is a suitable advantage. According to the results, BM25 underperformed against TF-IDF algorithm. This happened because parameters have not been tuned according to the dataset. Esparza et al. (2012) also demonstrated that ranking functions perform better than collaborative filtering algorithms.

The main conclusions from this literature survey are:

1. Ranking functions tend to perform better than well-known recommendation system algorithms like LDA and collaborative filtering.
2. Unstructured data provides certain amount of noise which is helpful for ranking function parameters.
3. Effects of length normalization on recommendation system have not been studied.

4 Penpal Recommendation System and Dataset Preparation

Penpal (online friends) recommendation system was set up from the online users. 630 users were asked about their interests, likes and relationships. Few volunteers helped in assessing in the matching of the penpals. Since every respondent was exclusively assigned one penpal, thus it gave 315 pairs as a result.

The response of the user was concatenated into a single sample or *document*. Hence the recommendation system situation was there for textual data for 630 users. MeTA (Massung et al.,

2016) toolkit was employed in the codebase of this project. Preprocessing on the data was done with the stopwords removal, tokenization and lemmatization. The dataset was divided into two parts:

- **Training set** : Text data of 504 samples (252 pairs)
- **Testing set** : Text data of 126 samples (63 pairs)

Average document length of the training set is 131 words. The observations followed that assessors paired up users whose response contained fewer words and similar interests (Table 1). Similar process was carried out for users having lengthy responses and similar interests. This may be due to the reason that user who is writing lesser words in the response form is less interested in having a penpal. Thus text similarity and length similarity plays a major role in penpal recommendation.

Description	Number of Pairs
Total Pairs in Training Dataset	252
Pairs matched with both document's length less than the average document length	62
Pairs matched with both document's length than the average document length	131
Others matches	53

Table 1: Document length characteristics in the training dataset

5 Feature Engineering

Here, the function would be modeled using Richard's curve. The curve is defined by,

$$g(x) = l + \frac{u - l}{(A + e^{-B(x-M)})^{1/\nu}} \quad (7)$$

Here, l is lower limit, u is upper limit and M, ν, A, B are free parameters

The distance metrics are taken to be just real numbers. Thus, $d(\mathbf{x}) = x$ and $d(\mathbf{y}) = y$. The trough of the curve should occur when, $d(\mathbf{x}) = d(\mathbf{y})$ or $x = y$. The function can be partitioned into three parts:

- monotonically decreasing Richard's curve, $g(x)$ when $x < y$
- monotonically increasing Richard's curve, $g(x)$ when $x > y$
- Value of heuristic function, $h(x, y) = 1$ when $x = y$

The definition of the function now looks like:

$$h(x, y) = \begin{cases} g(x_1) & \text{if } x < y \\ 1 & \text{if } x = y \\ g(x_2) & \text{if } x > y \end{cases} \quad (8)$$

where $x_1 \in [0, y]$, $x_2 \in [y, \infty]$ and $x_1, x_2 \subset x$
Applying equation (1), the constraint obtained

is:

$$\left. \frac{\partial h(x, y)}{\partial x} \right|_{x=y} = 0$$

From equation (2), the constraint obtained is:

$$\frac{\partial g(x_1)}{\partial x_1} < 0$$

From equation (3), the constraint obtained is:

$$\frac{\partial g(x_2)}{\partial x_2} > 0$$

From limiting condition (4), the constraint obtained is:

$$\lim_{x_1 \rightarrow 0} g(x_1) = b_1$$

$$\left. \frac{\partial g(x_1)}{\partial x_1} \right|_{x_1=0} \approx 0$$

From the limiting condition (5), the constraint obtained is:

$$\lim_{x_2 \rightarrow \infty} g(x_2) = b_2$$

$$\left. \frac{\partial g(x_2)}{\partial x_2} \right|_{x_2=\infty} \approx 0$$

From the limiting condition (6), the constraint obtained are:

$$\left. \frac{\partial g(x_1)}{\partial x_1} \right|_{x_1=y} \approx \left. \frac{\partial g(x_2)}{\partial x_2} \right|_{x_2=y} \approx 0$$

$$\lim_{x_1 \rightarrow y} g(x_1) = \lim_{x_2 \rightarrow y} g(x_2) = 1$$

After solving the constraints, and using the binomial approximation for small ν , the solutions obtained are:

$$h(x, y) = \begin{cases} 1 + \frac{b_1 - 1}{1 + e^{B_1(x - cy)}} & \text{if } x < y \\ 1 & \text{if } x = y \\ 1 + \frac{b_2 - 1}{1 + e^{-B_2(x - (1+c)y)}} & \text{if } x > y \end{cases} \quad (9)$$

where,

- B_1 and B_2 are growth parameters, can be typically set to 1
- c is the trough curvature, where $c \in (0, 1)$

According to the requirements, length similarity would be rewarded between query and document lengths. In that case, $x = |d|$ and $y = |q|$. Intuitively, that means the value of the scoring function will be higher if $|d| \approx |q|$.

Substituting the values, we get:

$$h(|d|, |q|) = \begin{cases} 1 + \frac{b_1 - 1}{1 + e^{B_1(|d| - c|q|)}} & \text{if } |d| < |q| \\ 1 & \text{if } |d| = |q| \\ 1 + \frac{b_2 - 1}{1 + e^{-B_2(|d| - (1+c)|q|)}} & \text{if } |d| > |q| \end{cases} \quad (10)$$

The nature of this feature can be plotted as shown against various parameters (Figure 2). It can be seen that it closely resembles the desired graph.

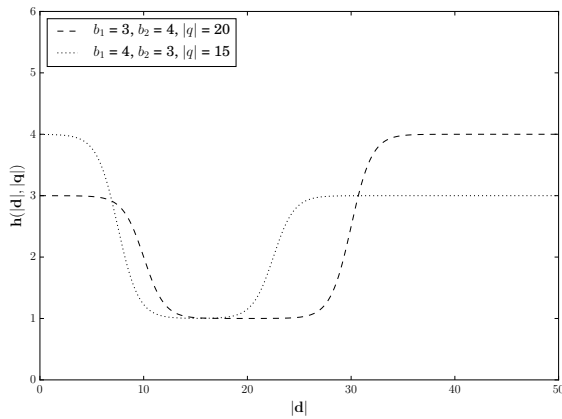


Figure 2: The nature of $h(|d|, |q|)$ with $B_1 = 1$, $B_2 = 1$ and $c = 0.5$

This feature can now be substituted in the ranking function in place of length normalization function. For example, the original BM25 function is:

$$score(q, d) = \sum_{t \in d \cap q} f(t, q) \log \frac{M + 1}{df(t)} \times \frac{(k + 1)f(t, d)}{f(t, d) + k \left(1 - b + b \frac{|d|}{avgdl}\right)}$$

The normalization feature $\left(1 - b + b \frac{|d|}{avgdl}\right)$ in the BM25 formula can be replaced with $h(|d|, |q|)$ to get our desired ranking function:

$$score(q, d) = \sum_{t \in d \cap q} f(t, q) \log \frac{M + 1}{df(t)} \times \frac{(k + 1)f(t, d)}{f(t, d) + k (h(|d|, |q|))} \quad (11)$$

6 Results and Conclusions

The relevant feedback of the dataset is limited. The human assessors have provided only one relevant document for one query. Thus, only MRR (Mean Reciprocal Rank) is used to assess the models.

$$MRR = \frac{1}{s} \sum_{i=1}^s \frac{1}{rank_i} \quad (12)$$

Here s is the size of the dataset and $rank_i$ is the position of the rank for the first relevant document of the i^{th} query.

The parameters of the baseline algorithms were tuned according to the training dataset and tested on the test set using MRR.

Model	Training Set MRR	Testing Set MRR
Length Similarity Heuristic with BM25	0.34	0.29
BM25	0.24	0.19
Pivoted Length Normalization	0.22	0.18
MPtf2ln	0.21	0.18
MDtf2ln	0.19	0.16
PL2	0.17	0.15
Dirichlet Prior	0.16	0.13

Table 2: Mean reciprocal rank (MRR) values on the data-set using various retrieval methods

The obtained MRR values have been given in (Table 2). The proximity feature has outperformed

by 52 % from regular BM25. It can be observed that query-document length similarity, not document length normalization has helped in this situation.

Parameter	Value
k (BM25 parameter)	2.8
b_1 (Left Bound)	2.9
b_2 (Right Bound)	3.7
B_1 (Growth parameter)	1
B_2 (Growth parameter)	1
c (Trough curvature)	0.5

Table 3: Parameters used in similarity feature in BM25

The parameters used in similarity feature (Table 3) with BM25 show that b_1 and b_2 are tuned around 3 and 4. This means that the magnitude of TF-IDF value is penalize around 3 or 4 times when the lengths are dissimilar.

Further work is to be carried out by using more evaluation tests, creating a better dataset, giving proofs for generalization and applying the algorithm into more applications like spelling correction.

References

- [Adomavicius and Tuzhilin2005] G. Adomavicius and A. Tuzhilin. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June.
- [Amati and Van Rijsbergen2002] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October.
- [Bellogín et al.2010] Alejandro Bellogín, Iván Cantador, and Pablo Castells. 2010. A study of heterogeneity in recommendations for a social music service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec ’10, pages 1–8, New York, NY, USA. ACM.
- [Cantador et al.2010] Iván Cantador, Alejandro Bellogín, and David Vallet. 2010. Content-based recommendation in social tagging systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys ’10, pages 237–240, New York, NY, USA. ACM.
- [Esparza et al.2012] Sandra Garcia Esparza, Michael P. OMahony, and Barry Smyth. 2012. Mining the real-time web: A novel approach to product recommendation. *Knowledge-Based Systems*, 29:3 – 11. Artificial Intelligence 2010.
- [Fang et al.2011] Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic evaluation of information retrieval models. *ACM Trans. Inf. Syst.*, 29(2):7:1–7:42, April.
- [Kwon and Kim2010] Joonhee Kwon and Sungrim Kim. 2010. Friend recommendation method using physical and social context.
- [Manning et al.2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- [Massung et al.2016] Sean Massung, Chase Geigle, and ChengXiang Zhai. 2016. Meta: A unified toolkit for text retrieval and analysis. *ACL 2016*, page 91.
- [Nishioka and Scherp2016] C. Nishioka and A. Scherp. 2016. Profiling vs. Time vs. Content: What does Matter for Top-k Publication Recommendation based on Twitter Profiles? - An Extended Technical Report. *ArXiv e-prints*, March.
- [Robertson and Walker1994] S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.
- [Singhal et al.1996] Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’96, pages 21–29, New York, NY, USA. ACM.
- [Suchal and Návrat2010] Ján Suchal and Pavol Návrat. 2010. *Full Text Search Engine as Scalable k-Nearest Neighbor Recommendation System*, pages 165–173. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Totti et al.2016] Luam C. Totti, Prasenjit Mitra, Mourad Ouzzani, and Mohammed J. Zaki. 2016. A query-oriented approach for relevance in citation networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW ’16 Companion, pages 401–406, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- [Wang and Yuan2009] Xin Wang and Fang Yuan. 2009. Course recommendation by improving bm25 to identity students’ different levels of interests in courses. In *New Trends in Information and Service Science, 2009. NISS’09. International Conference on*, pages 1372–1377. IEEE.

- [Wu et al.2008] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3):13:1–13:37, June.
- [Zhai and Lafferty2004] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April.
- [Zhai and Massung2016] ChengXiang Zhai and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool.